DOCUMENT RESUME

ED 468 073 TM 034 388

AUTHOR	Lee, Jaekyung; McIntire, Walter G.
TITLE	Using National and State Assessments To Evaluate the Performance of State Education Systems: Learning from the Cases of Kentucky and Maine. Research Report. Statewide Systemic Initiatives (SSI) Study.
INSTITUTION	Maine Univ., Orono.
SPONS AGENCY	National Science Foundation, Arlington, VA.
REPORT NO	SSI-RR-1
PUB DATE	2002-07-00
NOTE	31p.; For Research Report Number 2, see TM 034 389.
CONTRACT	REC-9970853
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	Academic Achievement; Case Studies; Elementary Secondary Education; *State Programs; Test Results; *Test Use; *Testing Programs
IDENTIFIERS	Kentucky Instructional Results Information System; Maine Educational Assessment; National Assessment of Educational Progress; Statewide Systemic Initiative

ABSTRACT

This study examined two major questions: do national and state assessments provide consistent information on the performance of state education systems, and what accounts for discrepancies between national and state assessment results if they are found? Data came from national and state assessments in grade 4 and grade 8 mathematics from 1992 to 1996 in Maine and Kentucky from the National Assessment of Educational Progress (NAEP), the Kentucky Instructional Results Information System (KIRIS), and the Maine Educational Assessment (MEA). NAEP and state assessments reported inconsistent results on the performance level of students in Maine and Kentucky across grades and years. Both MEA and KIRIS appear to have more rigorous performance standards, which reduces the percentage of students identified as performing at Proficient/Advanced level. These discrepancies may be understood in light of the differences between the NAEP and state assessments in their definitions of performance standards and the methods of standard setting. The size of achievement gaps between different groups of students appeared somewhat smaller on state assessments than on the NAEP, and the sizes of achievement gains from the state's ownassessments were considerably greater than that of the NAEP. These findings raise cautions in using either national or state assessments alone to evaluate the performance of particular state education systems. Policymakers and educators should become more aware of the unique features and limitations of the current national and state assessments. (Contains 3 figures, 17 tables, and 22 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 468 073

Research Report No. 1
Statewide Systemic Initiatives (SSI) Study

Using National and State Assessments to Evaluate the Performance of State Education Systems: Learning from the Cases of Kentucky and Maine

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
J. Lee
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

Jaekyung Lee, Ph.D.
Walter G. McIntire, Ph.D.
University of Maine

July 2002

TM034388

Sponsored by the National Science Foundation

2
BEST COPY AVAILABLE

Research Report No. 1
Statewide Systemic Initiatives (SSI) Study

Using National and State Assessments to Evaluate the Performance of State Education Systems: Learning from the Cases of Kentucky and Maine

by

Jaekyung Lee, Ph.D.
Walter G. McIntire, Ph.D.
University of Maine

Prepared for
Division of Research, Evaluation and Communication
Directorate for Education and Human Resources
National Science Foundation
Arlington, VA

July 2002

NSF Grant No. REC-9970853

Prepared by the University of Maine, Orono, Maine,
for the Division of Research, Evaluation and Communication,
Directorate for Education and Human Resources,
Bernice Anderson, Program Officer

July 2002

The conduct of this study and prepa. ıtion of this report was sponsored by the National Science Foundation, Directorate for Education and Human Resources, Division of Research, Evaluation and Communication, under Grant No. REC-9970853. Any opinions, findings, and conclusions or recommendations expressed in this report are those of the authors and do not necessarily represent the views of the National Science Foundation. This report and other related publications are available on the World Wide Web:
www.ume.maine.edu/naep/SSI

Table of Contents

Preface iv

Summary v

I. Research Objectives 1

II. Research Methods and Findings 2

How Do Students Measure Up Against National and State Performance Standards? 2

Differences in the Definition of Performance Standards 6

Differences in Standard-Setting (Identification of Cut Scores) Method 8

How Do Student Achievement Gaps Appear on National and State Assessments? 8

Differences in Testing Sample 10

Differences in Test Difficulty 11

How Much Has Student Performance Improved on National and State Assessments? 12

Differences in Test Changes and Equating 15

Differences in Test Stakes 18

III. Discussion 20

References 23

Preface

Evaluation of systemic school reform requires a systemic approach to data collection and analysis. The National Science Foundation's Statewide Systemic Initiatives (SSI), comprehensive state policies aimed at broad student populations, consider the effects of change on the total system over a sufficient period of time, and thus are distinctive in terms of the scale and nature of programs. We need to identify and fill the gaps between currently available data and methods and desired ones in assessing and understanding the performance of SSI states. We selected two SSI states, Kentucky and Maine, to explore two research questions:

First, what information is available on the academic performance of state education systems? While there are several ways to measure academic performance, we chose to focus on student achievement in mathematics. We examined whether and how the current national and state assessments can be used, together, to inform us of statewide academic performance. We also examined national and state assessments to determine if they produce inconsistent results and to explore reasons. Second, what methodological challenges are posed by multiple measures such as national, state, and local assessments as we seek to evaluate student and school performance? We attempted to identify appropriate methods for analyzing multi-dimensional achievement data: multiple measures of achievement collected through multiple types of assessments in the multiple levels of school system at multiple time points.

Research Report No. 1 is the product of our first-year SSI research study project, "Exploring Data and Methods to Assess and Understand the Performance of SSI States." During our first project year, we have focused on the first research question and produced significant findings. This first-year study examined the consistency of the National Assessment of Educational Progress (NAEP) and state assessments as statewide educational performance measures. Two states, Kentucky and Maine, were chosen for the study, and their students' 4th and 8th grade mathematics achievement data during the 1992-96 period were examined. Similarities and discrepancies between the NAEP and state assessments were examined in terms of three major statewide performance indicators that they produce: students' performance level, achievement gap, and achievement gain.

All of the research in this report was conducted by Dr. Jaekyung Lee (PI) and Dr. Walter McIntire (Co-PI). We are very grateful to the National Science Foundation for its financial support and to the University of Maine College of Education and Human Development for its administrative support. We acknowledge that both Maine and Kentucky state education agencies provided essential help by sharing their states' student assessment data and reports. We emphasize that the views expressed herein are solely those of the authors. Our special thanks go to Dr. Bernice Anderson at the National Science Foundation, Dr. Benjamin Wright, and Dr. Kenneth Wong who provided guidance and feedback throughout our project. We also thank Yuhong Sun, Jacqueline Henderson, Mary Anne Royal, and Amy Cates at the University of Maine, who provided research assistance and/or editorial assistance.

Summary

This study examined two major questions. Do national and state assessments provide consistent information on the performance of state education systems? What accounts for discrepancies between national and state assessment results if they are found?

Data came from national and state assessments in grade 4 and grade 8 mathematics from 1992 to 1996 in Maine and Kentucky: National Assessment of Educational Progress (NAEP), Kentucky Instructional Results Information System (KIRIS), and Maine Educational Assessment (MEA). Here is a very brief summary of major research findings:

1. NAEP and state assessments reported inconsistent results on the performance level of students in Maine and Kentucky across grades and years. Both MEA and KIRIS appear to have more rigorous performance standards, which reduces the percentage of students identified as performing at Proficient/Advanced level. These discrepancies may be understood in light of the differences between the NAEP and state assessments in their definitions of performance standards and the methods of standard setting.

2. The size of achievement gaps between different groups of students appeared somewhat smaller on state assessments than on the NAEP. The discrepancies may be explained by examining the differences between NAEP and state assessments in the representation of different student groups in their testing samples, the distribution of item difficulties in their tests, and differential impact of state assessment on low-performing students/schools.

3. The sizes of achievement gains from the states' own assessments were considerably greater than that of NAEP's. At the same time, the amount of difference is not always consistent across grades. These gaps and inconsistencies might be related to differences between the national and state assessments in the stakes of testing for school systems and changes in test format that impact test equating.

The study findings raise cautions in using either national or state assessment results alone to evaluate the performance of particular state education systems. This report also provides some preliminary analyses of the sources of inconsistencies and discrepancies between national and state assessments. Although these findings may not be generalized to all states, they suggest that policymakers and educators become more aware of the unique features and limitations of current national and state assessments. While the NAEP assessment can be used to cross-check and validate the states' own assessment results, each state's unique assessment characteristics (both policy and technical aspects) need to be considered. The study gives us implications for comparing and/or combining the results from national and state assessments.

I. Research Objectives

Since 1991, the National Science Foundation (NSF) has signed cooperative agreements with 26 states to undertake ambitious and comprehensive initiatives to reform science, mathematics, and technology education. This effort to improve public education is known as the Statewide Systemic Initiatives (SSI). While one of the NSF's drivers for systemic reform required improvement in the achievement of all students, the SSI program also explicitly requested that participating states seek ways to ensure that their systemic initiatives addressed equity issues.

Given statewide systemic reform efforts for academic excellence and equity, we need to know what information is available on the performance of state education systems. While the National Assessment of Educational Progress (NAEP) and individual state student assessments have been used to inform us of state-level performance, problems exist. On one hand, states are having difficulty in realigning their student assessment systems and tracking student achievement (CPRE, 1995). Moreover, most states use their statewide assessments for several purposes, some of which are incompatible (Bond, Braskamp, & Roeber, 1996). On the other hand, the NAEP state assessments provide highly comparable information on student achievement across the states, but they are not specifically aligned with the policies and standards of any given state. Thus, we need to examine whether and how the current NAEP and states' own student assessments can be used to inform us of systemwide academic performance. We also need to examine if the national and state assessments produce consistent results on the proficiency levels of students, the achievement gaps among different groups of students and their academic progress.

Our study is based on the premise that one must use multiple measures if the measures are to be used for evaluation that will result in consequences for students and/or their school systems (see AERA, APA, & NCME, 1999). Using more than a single measure may enhance the validity and fairness of evaluation. Nevertheless, it is really challenging to compare and link the results from national and state assessments which share some common technical features as a large-scale student assessment tool, but remain different in many other ways (NRC, 1999). If we simply focus on assessment results and compare them without looking into the assessments themselves, we are likely to draw erroneous conclusions. Once we make sure that the assessments are appropriate and comparable, then we must determine how to analyze the results which might be similar in some aspects and different in the others. One may be tempted to combine the results from two assessments by simply averaging them. But this approach can yield biased evaluations without considering each assessment's unique features (i.e., goals, content, process, context, consequences) and technical qualities. We need to identify factors that produces discrepancies and make evaluation conditional upon those factors.

In light of these concerns, we conducted a systematic analysis of currently available statewide student assessment data ,using NAEP and state assessments. We addressed the consistency of these assessments for producing information on the performance of states. The objective of this study was to identify and explain the gaps between national and state assessments in light of three major educational system performance indicators: (1) students' performance level, (2) achievement gap, and (3) achievement gain. We also explored some of the factors that might explain any discrepancies in the NAEP and state assessment results.

II. Research Methods and Findings

We selected and examined two SSI states, Kentucky and Maine, which (1) put student assessment systems in place early enough to gather baseline data and monitor their progress, (2) made their assessments more in line with the goals of their education reform initiatives than other states, and (3) adopted similar performance standards to those in the NAEP. We utilized data collected from the states' student assessments, that is, Kentucky Instructional Results Information System (KIRIS) and Maine Educational Assessment (MEA) in mathematics at grade 4 and grade 8 from 1992 through 1996. We also used the NAEP state assessment data for cross-check and cross-state comparisons: the NAEP state mathematics assessments in both Maine and Kentucky were collected for 4th and 8th graders in 1992 and 1996. The NAEP state assessment was administered to a random sample of each state's fourth and eighth graders while both MEA and KIRIS were given to the virtually entire populations of Maine and Kentucky fourth and eighth graders. Our data do not include students which were exempt or absent from testing and whose test scores were not reported or missing for any reasons.

Several concerns have been raised about what data is required to adequately assess the performance of a system (Laguarda et al., 1994). Do the tests exist? If so, are they aligned with the curriculum content promoted by national and state education goals? Are the results available in a form compatible with national and state performance standards? Have the assessments been equated across the years and grade levels to track performance gains? Assessments in our study states meet the above-mentioned criteria, but it remains to be seen whether these state assessments produce the same information as the NAEP regarding the performance of the systems as a whole. We not only conducted analysis of the raw data but also reviewed information available from existing technical reports or manuals on the NAEP and state assessments. In the following sections, three major aspects of educational system performance are examined: the level of student achievement, the size of the student achievement gap, and the amount of achievement gain.

How Do Students Measure Up Against National and State Performance Standards?

Previous comparisons of national and state assessment results have shown that the percentages of students reaching the proficient level on NAEP are generally lower than on the state assessments. These results have been interpreted by educational policymakers as implying that for many states, NAEP proficiency levels are more challenging than the states' own and that state standards are still not high enough (see National Education Goals Panel, 1996). However, differences between NAEP and state assessments in the purpose of their performance standards were also noted and their comparability was questioned (Linn, 2000). The issue of comparability is much less problematic in the cases of Maine and Kentucky assessments, because they modeled their frameworks closely after NAEP and adopted very challenging performance standards.

The NAEP achievement levels, as authorized by the NAEP legislation and adopted by the National Assessment Governing Board (NAGB), are collective judgments, gathered from a broadly representative panel of teachers, education specialists, and members of the general public, about what students should know and be able to do relative to a body of content reflected in the NAEP assessment frameworks. For reporting purposes, the achievement level cut scores for each grade are placed on the traditional NAEP scale resulting in four ranges: Below Basic, Basic, Proficient, and Advanced.

Both Maine and Kentucky have achievement levels that are very similar to the NAEP levels. In Maine, proficiency levels were introduced into the MEAs in 1995, and students were identified as being in Novice, Basic, Advanced, or Distinguished levels of achievement. In Kentucky, four corresponding categories were established for the KIRIS in 1992: Novice, Apprentice, Proficient, and Distinguished. While Kentucky set its student performance goal at the Proficient level on the KIRIS as a result of statewide education reform (i.e., 100% of students proficient in 20 years), Maine did not specifically link their performance standards with the MEA proficiency levels. Despite the lack of a standards-assessment linkage, it was reasonable to say that Maine also set its performance expectation for all students to the level of being "Advanced" on the MEA. Category labels and brief generic definitions are shown in Table 1.

Table 1. Comparison of NAEP, KIRIS and MEA Definitions of Student Performance Levels

NAEP	KIRIS	MEA
Below Basic Students have little or no mastery of knowledge and skills necessary to perform work at each grade level.	Novice The student is beginning to show an understanding of new information or skills.	Novice Maine students display partial command of essential knowledge and skills.
Basic Students have partial mastery of knowledge and skills fundamental for proficient work.	Apprentice The student has gained more understanding, can do some important parts of the task.	Basic Maine students demonstrate a command of essential knowledge and skills with partial success on tasks involving higher-level concepts, including application of skills.
Proficient Students demonstrate competency over challenging subject matter and are well prepared for the next level of schooling.	Proficient The student understands the major concepts, can do almost all of the task, and can communicate concepts clearly.	Advanced Maine students successfully apply a wealth of knowledge and skills to independently develop new understanding and solutions to problems and tasks.
Advanced Student show superior performance beyond the proficient grade-level mastery.	Distinguished The student has deep understanding of the concept or process and can complete all important parts of the task. The student can communicate well, think concretely and abstractly, and analyze and interpret data.	Distinguished Maine students demonstrate in-depth understanding of information and concepts.

BEST COPY AVAILABLE

In order to see how students in Kentucky and Maine meet national and state performance standards, we compared NAEP and state math assessment results on student performance in 1992 and 1996 (1996 only for Maine because the MEA lacked performance standards in 1992). As shown in Table 2, the percentage of students at or above the NAEP Proficient level is smaller than at or above the MEA Advanced level. Specifically, the difference is remarkable at grade 8: 31% of Maine eighth grade students meet the NAEP's Proficient level in math as of 1996, whereas only 9% of the students meet the MEA's Advanced level. Thus, as Maine sticks more to the state's own performance goals, it ends up with a longer way to go. On the other hand, the definition of Basic performance level seems to be more convergent between the NAEP and MEA. Whether we base our judgment of Maine students' performance on the NAEP or MEA achievement levels, we come to the same conclusion that approximately one fourth of the student population in Maine does perform below the Basic level across grades and subjects examined.

Table 2. Percentages of Maine 4th and 8th Graders by Performance Level on 1996 NAEP and MEA Mathematics

NAEP		MEA	
Grade 4			
Advanced	3	Distinguished	8
Proficient	24	Advanced	15
Basic	48	Basic	55
Below Basic	25	Novice	22
Grade 8			
Advanced	6	Distinguished	1
Proficient	25	Advanced	8
Basic	46	Basic	62
Below Basic	23	Novice	29

On the other hand, comparison of NAEP and KIRIS assessment results reveal more inconsistent performance patterns. Table 3 shows the results of 1992 assessments in which the percentage of students below the NAEP Basic level is smaller than the KIRIS Novice level, whereas the percentage of students at or above the NAEP and KIRIS Proficient level is more congruent. However, the results of the 1996 assessments reversed the pattern: the percentage of students below the NAEP Basic level is greater than the KIRIS Novice level (see Table 4).

Table 3. Percentages of Kentucky 4th and 8th Graders by Performance Level on 1992 NAEP and KIRIS Mathematics

NAEP		KIRIS	
Grade 4			
Advanced	1	Distinguished	2
Proficient	12	Proficient	3
Basic	38	Apprentice	31
Below Basic	49	Novice	65
Grade 8			
Advanced	2	Distinguished	3
Proficient	12	Proficient	10
Basic	37	Apprentice	24
Below Basic	49	Novice	63

Table 4. Percentages of Kentucky 4th and 8th Graders by Performance Level on 1996 NAEP and KIRIS Mathematics

NAEP		KIRIS	
Grade 4			
Advanced	1	Distinguished	5
Proficient	15	Proficient	9
Basic	44	Apprentice	56
Below Basic	40	Novice	30
Grade 8			
Advanced	1	Distinguished	12
Proficient	15	Proficient	16
Basic	40	Apprentice	36
Below Basic	44	Novice	36

By and large, the performance standards for the KIRIS and MEA appear to have been set at comparable or even higher levels than the standards for NAEP: the percentage of students at or above the NAEP Proficient level is equal to or smaller than at or above the KIRIS Proficient level and MEA Advanced level. Nevertheless, the comparison of the NAEP, MEA and KIRIS assessment results identified inconsistent percentages of students in their corresponding performance categories. In the following sections, we explored potential factors that might explain those gaps or inconsistencies in standards-based performance results by examining how the definition of performance standards and standard-setting method differed between the national and state assessments.

Differences in the Definition of Performance Standards

As shown above, NAEP, Kentucky and Maine assessments all employed four performance standards or achievement levels. It appears that each tried to keep the standards to a reasonable number, avoiding potential problems with too few (no recognition of modest progress) or too many standards (inaccuracy of classification). Further, the KIRIS technical manual (1995) describes the difficulty that Kentucky faced in naming performance standards, particularly choosing the term "proficient" for the level of success:

> Its only drawback was that NAEP uses that term; since KIRIS will be linked to NAEP, and because NAEP's standard of "proficient" likely will be at least somewhat different from Kentucky's, there was concern about confusion between the two. However, all things considered, "Proficient" was judged to be the most appropriate term. (p. 65)

However, the real issue is operational definitions. The definition of standards affects the level of cut scores associated with the standards (Jaeger & Mills, 2001). Part of the differences between NAEP and state performance results can be explained by comparing performance level definitions by subject and grade. NAEP has both grade-specific and subject-specific definitions of performance levels, while the MEA has only subject-specific definitions and KIRIS lacks both subject-specific and grade-specific standards. Particularly the KIRIS performance standards were criticized for their vagueness (Hambleton et al., 1995). The presence or absence of clearly-stated and well-specified definitions of performance standards and achievement levels by grade and subject may help explain the differences in outcomes.

Table 5 provides definitions of MEA and NAEP math achievement levels; the 4th grade-specific definition is shown for NAEP while an across-grade definition is shown for the MEA. It is obvious that the NAEP has more clear and specific definitions with performance indicators than does the MEA. Definitions of "Basic" look very similar in that both assessments require demonstrations of student ability to solve some simple, routine problems with limited reasoning and communication. In contrast, the MEA definition of "Advanced" appears somewhat more rigorous than the NAEP definition of "Proficient": the former requires the student to solve both routine and non-routine (many) problems with effective reasoning and communication, whereas the latter requires the student to consistently solve routine problems (as distinct from complex, nonroutine problems) with successful reasoning and communication. However, both the complexity and non-routineness of any math problem is a matter of degree and subject to personal judgement. Consequently, without careful elaboration of standards by subject and grade, it is very unlikely that we will find congruence between national and state assessments in the percentages of students at the proficiency levels even with similar generic definitions and labels.

Table 5. Comparison of NAEP and MEA Definition of Mathematics Performance Levels

	NAEP (Grade 4-Specific)		MEA (Grade-Free)	
Below Basic		Novice. Maine students demonstrate some success with computational skills, but have great difficulty applying those skills to problem-solving situations. Mathematical reasoning and communication skills are minimal.		
Basic. Fourth-grade students should show some evidence of understanding the mathematical concepts and procedures in the five NAEP content strands. Estimate and use basic facts to perform simple computations with whole numbers; show some understanding of fractions and decimals; and solve some simple real-world problems; use four-function calculators, rulers, and geometric shapes (though not always accurately). Their written responses are often minimal and presented without supporting information.		Basic. Maine students can solve routine problems, but are challenged to develop appropriate strategies for non-routine problems. Solutions sometimes lack accuracy; reasoning and communications are sometimes limited.		
Proficient. Fourth-grade students should consistently apply integrated procedural knowledge and conceptual understanding to problem solving in the five NAEP content strands. Use whole numbers to estimate, compute, and determine whether results are reasonable; have a conceptual understanding of fractions and decimals; solve real-world problems; use four-function calculators, rulers, and geometric shapes appropriately; employ problem-solving strategies such as identifying and using appropriate information. Their written solutions are organized and presented both with supporting information and explanations of how they were achieved.		Advanced. Maine students solve routine and many non-routine problems and determine the reasonableness of the solutions using estimation, patterns and relationships, connections among mathematical concepts, and effective organization of data. These students make important connections of mathematics to real-world situations, do accurate work, and communicate mathematical strategies effectively.		
Advanced. Fourth-grade students should apply integrated procedural knowledge and conceptual understanding to complex and nonroutine real-world problems in the five NAEP content strands. Solve complex and non-routine real-world problems; display mastery in the use of four-function calculators, rulers, and geometric shapes; draw logical conclusions and justify answers and solution process; go beyond the obvious in their interpretations and be able to communicate their thoughts clearly and concisely.		Distinguished. Maine students demonstrate an in-depth understanding of mathematics by applying sound reasoning to solve non-routine problems using efficient and sometimes innovative strategies. These students make connections among mathematical concepts and extend their understanding of specific problems to more global or parallel situations. They can communicate mathematically with effectiveness and sophistication		

Source. Figure 3.1 in Reese et al. (1997). *NAEP 1996 Math Report Card for the Nation and the States;* Maine Department of Education (1996). *MEA Performance Level Guide: Grade 4.*

BEST COPY AVAILABLE

Differences in Standard-Setting (Identification of Cut Scores) Method

The NAEP math achievement levels were set following the 1990 assessment and further refined following the 1992 assessment. In developing the threshold values (cut scores) for the levels, a panel of judges rated a grade-specific item pool using the policy definitions of the NAGB. The NAEP performance standard-setting process employed a variant of Angoff method (NCES, 1997). The judges (24 at grade 4 and 22 at grade 8) rated the questions in terms of the expected probability that a student at a borderline achievement level would answer the questions correctly (for multiple-choice and short constructed-response items) or receive scores of 1, 2, 3, and 4 for the extended constructed-response items. The results from the first round of approximation were adjusted by going through subsequent rounds of review/revision processes.

The 1992 math achievement levels were evaluated by several groups including the National Academy of Education. They raised serious concerns about the reliability and validity of the current achievement levels, concluding that the Angoff judgement method was not reasonable and could yield misleading interpretations (see Shepard et al., 1993; U.S. General Accounting Office, 1993). The MEA Performance Level Guide (1994-95) from Maine Department of Education also criticizes the NAEP standard-setting process as unrealistic and unreliable. It emphasizes the need for a different approach for the MEA in that the MEA employs a totally open-response format (scored on a 0-4 scale). Thus, the MEA standard-setting process utilized a totally different method which involved judges matching actual student work to the pre-determined definitions. By matching student work to the performance level definitions, ranges of the scale where cut-points are likely to be found were identified. Once the ranges were identified, judges examined large volumes of student work within the range and the cut points were identified based on the ratings of all judges.

The Kentucky standard-setting process shares some common features with Maine. First, Kentucky's standard setting was done on open-response items only; no multiple-choice items were included in the process. Second, standard setting was done by examining actual student work rather than by investigating test items. Third, standard setting was initiated as a result of standards-based statewide education reform and designed for monitoring systemwide progress toward the goal.

Studies show that different standard setting methods yield inconsistent results (Jaeger, 1989). In our case, it is not clear how the use of different standard-setting methods affected the cut scores and resulting estimation of the percentage of students at multiple achievement levels. The lack of comparability across different standard setting methods is further complicated by the use of different performance level definitions by NAEP and state assessments. Any effort to directly compare and/or combine NAEP and state assessments' performance level results may be misleading without considering these differences and their potential influences.

How Do Student Achievement Gaps Appear on National and State Assessments?

When the performance of a school system is evaluated from an equity perspective, the size of student achievement gap becomes an important indicator of the system performance. We examined whether the sizes of achievement gaps between different groups of students are consistent between the states' own assessments and the NAEP. We selected four major student background variables (i.e., gender, race, parental education, and Title I program participation) that are available both in the national and state assessments and computed standardized gap estimates (see Table 6 and Table 7). As the student achievement gaps reported in standard deviation units incorporate differences in test score distribution as scaling

artifacts, any discrepancies between the national and state assessments in the size of achievement gaps among the same student groups requires explanation.

By and large, the standardized gap estimates in standard deviation units turned out to be smaller on the state's own assessments than on the NAEP although their discrepancies were very modest. The only exception to this pattern was a gender gap in Kentucky 8th grade math where the gap appeared larger on NAEP than on KIRIS. Regardless of the type of assessment in both states, however, it needs to be noted that the score differences between male and female students are relatively very small (hardly different from zero) in comparison with racial, social or academic gaps. In Maine, the gap between students whose parents had a high school education or more and students whose parents had less than a high school education was as large as the gap between Title I students and non-Title I students. In Kentucky, the gap between white and minority students was also as large as the gap between Title I students and non-Title I students.

Table 6. Maine 8th Grade Math Achievement Gaps on 1996 MEA and NAEP by Gender, Parental Education, and Title I Participation

	Standardized Gap		
Assessment	Gender	Parental Education	Title I
MEA	0.01	0.74*	0.80*
NAEP	0.06	0.86*	0.92*

Note: Parental education gap is between students who reported having parents with high school or more education vs. less than high school. Standardized gap is obtained by dividing the scale score gap between two concerned groups by their pooled standard deviation. Asterisk indicates that the gap is statistically significant at the .05 level.

Table 7. Kentucky 8th Grade Math Achievement Gaps on 1996 KIRIS and NAEP by Gender, Race, and Title I Participation

	Standardized Gap		
Assessment	Gender	Race	Title I
KIRIS	0.09*	0.53*	0.53*
NAEP	0.01	0.60*	0.85*

Note: Race gap is between white students and minority students. Standardized gap is obtained by dividing the scale score gap between two concerned groups by their pooled standard deviation. Asterisk indicates that the gap is statistically significant at the .05 level.

Differences in Testing Sample

Why do the gaps among different groups of students appear slightly larger on the NAEP than on the state assessments? One factor to consider is whether the NAEP testing sample is equivalent to the state assessment testing sample. Because NAEP employed a multistage stratified random sampling method, its sample was designed to properly represent major racial/ethnic and socioeconomic groups of students in each participating state (with an expectation of relatively small-size groups like Asian-Americans). In contrast, the state assessments do not involve any kind of sampling to select examinees, and their available testing samples are supposed to fully represent all student groups across the state. The exceptions include students with learning disabilities and limited English proficiency for whom the national and state assessments did not use exactly the same inclusion criteria for their testing and reporting.

To determine if the student groups compared in the previous section have equal representations in NAEP vs. state testing samples, we compared the percentage of students broken down by gender, race, parental education, and Title I. For gender and parental education, both NAEP and state assessments show exactly the same distributions. For race, there is 2% difference (11% minority for KIRIS vs. 13% minority for NAEP) in Kentucky but they are virtually identical considering the NAEP percent estimate's standard error of 1.03. For Title I participation, we found a significant difference: there is a 7% difference in Maine and 10% difference in Kentucky (see Figure 1). While Maine data shows slightly higher percentage of Title I students in the MEA than in the NAEP sample, Kentucky data shows the opposite pattern. We don't know the reason for these differences in both states but it might be due to misidentification or sampling error with regard to Title I group. Any overrepresentation or underrepresentation of Title I students in the samples who are mostly low-performing might be related to the difference between the NAEP and state assessments in their estimation of the Title I achievement gap.

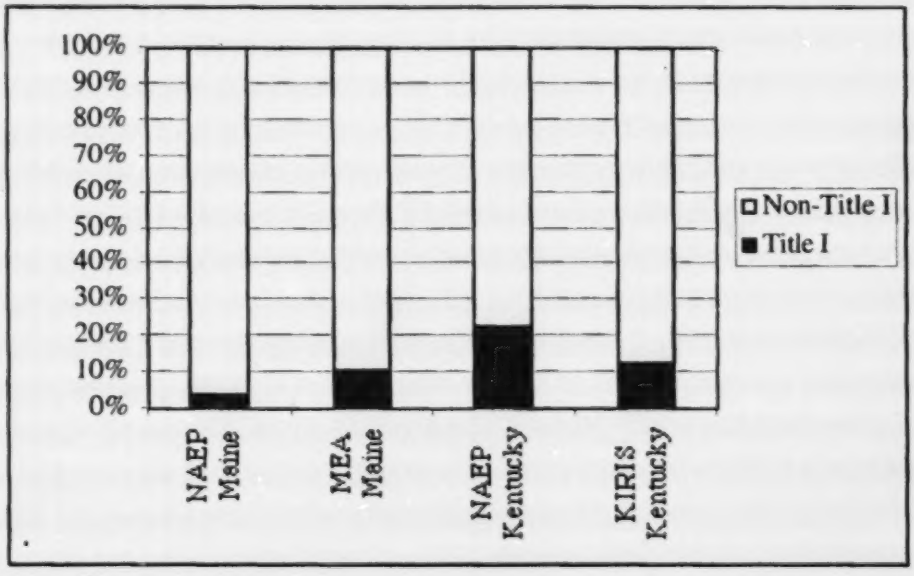

Figure 1. The Percentage of Title I vs. No-Title I Students in the 1996 NAEP, MEA, KIRIS 8th Grade Math Assessment Samples

Differences in Test Difficulty

Another potential factor that might influence the achievement gap estimates is test item difficulty. If some of the test items are more difficult for one group than for another group at the same level of proficiency, then it will affect the estimation of achievement gap. This can happen when the test items have an inherent bias or involve significant unequal opportunity to learn among different groups. Both the national and state assessments went through procedures to check against potential test bias and to conduct differential item functioning (DIF) analysis.

Assuming that all of the test items are equally difficult across different gender, race/ethnicity, and social groups , we need to consider how well those different assessments provide information on student achievement at different levels of proficiency. Although the assessments using more focused, challenging performance-type exams may provide richer information on the process of student learning (Neil et al., 1995), they may not serve all students equally well. Comparison of NAEP grade 8 mathematics test item information showed that the extended-response tasks provide much more information than both multiple-choice and short constructed-response items at the upper end of proficiency scale but less information at the lower end of the scale (see Dossey, Mullis, and Jones, 1993).

The NAEP employs more test items with a combination of multiple-choice and constructed-response items which produce wider range of item difficulties, whereas the state assessments with relatively limited number of only constructed-response items tend to have very narrow distributions of item difficulties (see Table 8 and Table 9). Lower item scores indicate greater difficulty, and both MEA and KIRIS may have been more difficult for low-performing students than the NAEP; most of the state test item scores are below .5. The MEA and KIRIS were likely to produce greater achievement gaps as they lacked test items that could measure student achievement at the lower end. Although our use of standardized gap measure takes into account potential difference in the score distributions, further investigation is needed.

Table 8. Maine Test Item (Easiness) Scores in 1996 MEA and NAEP 8th Grade Math

	Item Scores										
	.00-.10	.11-.20	.21-.30	.31-.40	.41-.50	.51-.60	.61-.70	.71-.80	.81-.90	.91-1.00	Total N
MEA	0	1	2	2	2	1	0	0	0	0	8
NAEP	0 (0)	4 (2)	20 (7)	18 (3)	14 (2)	18 (1)	17 (3)	21 (5)	25 (0)	10 (1)	147 (24)

Note. Only common items across test forms are available for the MEA. The number of entire MEA test items is 30 and all are polytomously-scored constructed-response items. Numbers in parenthesis indicate the number of polytomously-scored constructed-response items among all NAEP test items; the remainder includes multiple-choice items and dichotomously-scored constructed-response items. For dichotomoulsy-scored items (0, 1 scoring), the item score is the proportion of students who correctly answered each item. For polytomously-scored items, the item score is adjusted by dividing its mean by the maximum number of points possible.

Table 9. Kentucky Test Item (Easiness) Scores in 1996 KIRIS and NAEP 8th Grade Math

	Item Scores										
	.00-.10	.11-.20	.21-.30	.31-.40	.41-.50	.51-.60	.61-.70	.71-.80	.81-.90	.91-1.00	Total N
KIRIS	0	0	13	8	8	1	0	0	0	0	30
NAEP	1 (1)	15 (4)	21 (6)	24 (3)	17 (3)	19 (2)	17 (4)	13 (0)	15 (1)	5 (0)	147 (24)

Note. All of the above KIRIS items are polytomously-scored constructed-response items. Numbers in parenthesis indicate the number of polytomously-scored constructed-response items among all NAEP test items; the remainder includes multiple-choice items and dichotomously-scored constructed-response items. For dichotomoulsy-scored items (0, 1 scoring), the item score is the proportion of students who correctly answered each item. For polytomously-scored items, the item score is adjusted by dividing its mean by the maximum number of points possible.

The fact that state assessments in Maine and Kentucky were more challenging and difficult than their NAEP counterpart may reflect the two states' exceptionally high content and performance standards for all students. While the assessment by itself may be partly responsible for the discrepancy in the estimated size of student achievement gaps, we can think of the effect of broader assessment-driven state education policies and practices that might have functioned as achievement equalizers. Suppose that state assessment has a greater impact on lower-performing students and their schools which may pay more attention to the state test as an accountability measure and teach to the test. Student achievement scores on the state assessments may turn out to appear more equitable than on the NAEP. It remains to be investigated whether both states' assessment-driven school reform policies could have made any differential impact on schools at different performance levels and whether this could have made student achievement gaps appear smaller on the state assessments than on the NAEP.

How Much Has Student Performance Improved on National and State Assessments?

In the midst of standards-based school reform movement, every school system is expected to make continuous academic progress. The central question is whether the current NAEP and state assessments allow us to consistently keep track of system performance. To examine this issue, we first looked at changes in MEA and KIRIS student performance. Table 10 shows that the overall Maine performance trends in mathematics are highly positive across grade levels over the 1990-1997 period. Table 11 also shows that the overall Kentucky performance trends in mathematics are highly positive across grade levels over the 1992-1998 period. This successive cohort comparison method requires that the same grades of students are tested successively over time and their test scores are compared. The validity of this method for evaluating a school system's academic progress may be challenged if there are significant demographic changes in its student population over time and high level of student mobility during the school years. But we assumed that this potential problem is highly minimal at the aggregate state level.

Table 10. 1990-1997 MEA State Average Scale Score Trends in Mathematics

	1990	1991	1992	1993	1994	1995	1996	1997
Grade 4	255	265	270	270	285	285	330	320
Grade 8	300	305	305	315	325	325	350	360

Note. Scores were held constant in 1995 because of the change in test format.

Table 11. 1992-1998 KIRIS State Accountability Index Score Trends in Mathematics

	1992	1993	1994	1995	1996	1997	1998
Grade 4/5	17.8	22.3	34.2	41.8	38.9	44.8	44.4
Grade 7/8	23.8	22.8	31.4	48.9	47.3	53.8	51.4

Note. Math index is based upon the combination of on-demand and portfolio scores for 1993 and 1994 and on-demand scores only for 1995-1998.

Despite such positive performance trends based on the state assessment results, it is worthy to examine whether both Maine and Kentucky students made comparable amount of progress on the National Assessment of Educational Progress in mathematics. Earlier comparison of the KIRIS and NAEP achievement gains showed discrepancies (Hambleton et al., 1995). Using the NAEP and state assessment 4th and 8th grade math results in 1992 and 1996, we compared achievement gains from 1992 to 1996.

Tables 12 and 13 compare Maine student performance improvement levels based on the NAEP and MEA assessment results. Because NAEP and MEA scores employ different scales, a common metric in standard deviation units was established. Specifically, student standard deviations as obtained from the MEA 1996 mathematics assessment results were used to compute MEA standardized gain, while Maine's standard deviations from the 1996 NAEP state assessment results were used to compute NAEP standardized gain.

Table 12. Maine 4th Grade Math Score Gains on MEA and NAEP from 1992 to 1996

Assessment	1992	1996	Raw Gain	Standardized Gain
MEA	270	330	60*	0.39
NAEP	231	232	1	0.03

Note. Asterisk indicates that the gain is statistically significant at the .05 level.

Table 13. Maine 8th Grade Math Score Gains on MEA and NAEP from 1992 to 1996

Assessment	1992	1996	Raw Gain	Standardized Gain
MEA	305	350	45*	0.34
NAEP	279	284	5*	0.16

Note. Asterisk indicates that the gain is statistically significant at the .05 level.

Tables 14 and 15 compare Kentucky student performance improvement levels based on the NAEP and KIRIS assessment results. Because NAEP and KIRIS report gains in the percent of students meeting their own performance standards, a common metric in Cohen's h units was established. Specifically, percents of students at or above Proficient level as obtained from the KIRIS 1992 and 1996 assessment results were used to compute KIRIS standardized gain, while their counterparts from the 1992 and 1996 NAEP state assessment results were used to compute NAEP standardized gain.

Table 14. Kentucky 4th Grade Math Percent Proficient Gains on KIRIS and NAEP from 1992 to 1996

Assessment	1992	1996	Raw Gain	Standardized Gain
KIRIS	5	14	9*	0.32
NAEP	13	16	3	0.08

Note. Asterisk indicates that the gain is statistically significant at the .05 level.

Table 15. Kentucky 8th Grade Math Percent Proficient Gains on KIRIS and NAEP from 1992 to 1996

Assessment	1992	1996	Raw Gain	Standardized Gain
KIRIS	13	28	15*	0.38
NAEP	14	16	2	0.06

Note. Asterisk indicates that the gain is statistically significant at the .05 level.

As shown in Tables 12, 13, 14 and 15, we find overall statewide academic improvement in Maine and Kentucky since the early 1990s as measured by the MEA and KIRIS. However, the sizes of state math score gains tend to be somewhat greater than are observed in national assessment results (NAEP): ap-

proximately 13 times larger for grade 4 math, and twice as large for grade 8 math in the case of Maine; approximately 4 times larger for grade 4 math, and 6 times larger for grade 8 math in the case of Kentucky.

Both NAEP and state assessments face simultaneous goals of measuring trends in educational performance and providing information about student achievement on progressive curricular goals. NAEP uses several procedures to maintain the stability required for measuring trends, while still introducing innovations (Mullis et al., 1991). To keep pace with developments in assessment methodology and research about learning in each subject area, NAEP updates substantial proportions of the assessments with each successive administration. However, in some subject areas, NAEP conducts parallel assessments to provide separately for links to the past and the future. In the MEA and KIRIS, equating tests across years has been done by comparing any two adjacent years' test difficulties based on the items common to the tests both years. Nevertheless, drastic changes in the test content and format of tests raise doubts about whether their test equating is reliable and acceptable. In the following sections, we describe changes in the content and format of national and state assessments between 1992 and 1996, and explore how those changes might have affected results on test equating and performance gains.

Differences in Test Changes and Equating

Test specifications provide information on the content and format of national and state assessments. Table 16 shows the percentages of questions in 1992 and 1996 NAEP grade 4 and grade 8 math assessments. Questions could be classified under more than one content strand. It appears that changes were made in two content areas, "number sense, properties and operations" (fewer questions) and "algebra and functions" (more questions), which reportedly reflect the refinement of the NAEP math assessment to conform with recommendations from the NCTM standards (Reese at al., 1997).

Table 16. Percentage Distribution of NAEP Math Test Items by Content Strand and Grade

	Grade 4		Grade 8	
	1992	1996	1992	1996
Content Area				
Number Sense, Properties & Operation	45	40	30	25
Measurement	20	20	15	15
Geometry and Spatial Sense	15	15	20	20
Data Analysis, Statistics and Probability	10	10	15	15
Algebra & Functions	10	15	20	25
Total Percentage	100	100	100	100

Table 17. Percentage Distribution of KIRIS Math Test Items by Content Strand and Grade

	Grade 4		Grade 8	
	1992	1996	1992	1996
Content Area				
Number	13	14	20	16
Procedures	20	17	13	22
Space/Dimension	13	14	13	11
Measurement	13	14	20	16
Change	13	10	7	16
Structure	8	10	7	5
Data	20	21	20	14
Total Percentage	100	100	100	100

Source. Kentucky Department of Education (1995). *KIRIS Accountability Cycle 1 Technical Manual;* Kentucky Department of Education (1997). *KIRIS Accountability Cycle 2 Technical Manual.*

Reportedly, the curriculum and assessment frameworks for both the KIRIS and the MEA were based on those employed in creating NAEP tests. Table 17 shows the distribution of open-response KIRIS math items by year and grade across content areas. The entire KIRIS framework was consistent with the NAEP framework for mathematics. It appears that there were relatively large changes between 1992 and 1996 in KIRIS. Like NAEP, a single item in KIRIS often addresses more than one content area, which may have made the distribution of items less stable over time. The same can be said of the MEA.

While changes in test content tend to be minimal for both national and state math assessments, changes in test format and scoring standards also affect the stability of scores. The KIRIS, which started with a mix of performance exam items (i.e., writing portfolios, performance events, an on-demand essay, and open-response items) and multiple-choice items in 1992, later dropped multiple choice items. Liwewise, the MEA, which began as a combination of both multiple-choice and constructed-response questions, shifted to entirely constructed-response questions in 1995. The MEA 1994-1995 guide explains the rational for this change as follows:

> The findings of research studies are conclusive: heavy reliance on the multiple-choice format in high-stakes testing can have a negative effect on curriculum and instruction. On the other hand, the positive effect on curriculum and instruction associated with alternative modes of testing is widely recognized... MEA's use of "alternative" types of items is limited at this point to open-response items. Techniques for improving the data quality from portfolios and performance events for purpose of large-scale assessment are currently being investigated and refined. But the data quality from results of on-demand open-response testing, as used in Maine, is technically very sound. (p. 3)

BEST COPY AVAILABLE

Less dramatic but notable changes have been also made in the NAEP assessments. As a consequence of major revisions in the NAEP content framework in response to national standards, the 1990 NAEP assessment included a broad range of questions that required students to solve problems in both constructed-response and multiple-choice formats. For 1992, to increase NAEP's responsiveness to the then-published standards, the math assessment was nearly doubled in scope to provide greater emphasis on constructed-response questions and innovative problem-solving situations (Dossey, Mullis, and Jones, 1993). In 1996 NAEP testing, more than 50% of student assessment time was devoted to constructed-response questions.

Figure 2 illustrates these changes. While both national and state assessments shifted from multiple-choice items (MC) to more constructed-response questions (CR) including extended constructed-response questions that required students to provide an answer and a corresponding explanation, the extent of changes was greater in state assessments than NAEP between 1992 and 1996. If test score tends to drop right after introduction of a new test form (Linn et al., 1990), we might expect relatively smaller achievement gains on state assessment that changed its test format more substantially. But the pattern of actual achievement gains on NAEP, MEA, and KIRIS does not meet this expectation and asks for further examination of other factors that might have overridden the effect of test changes.

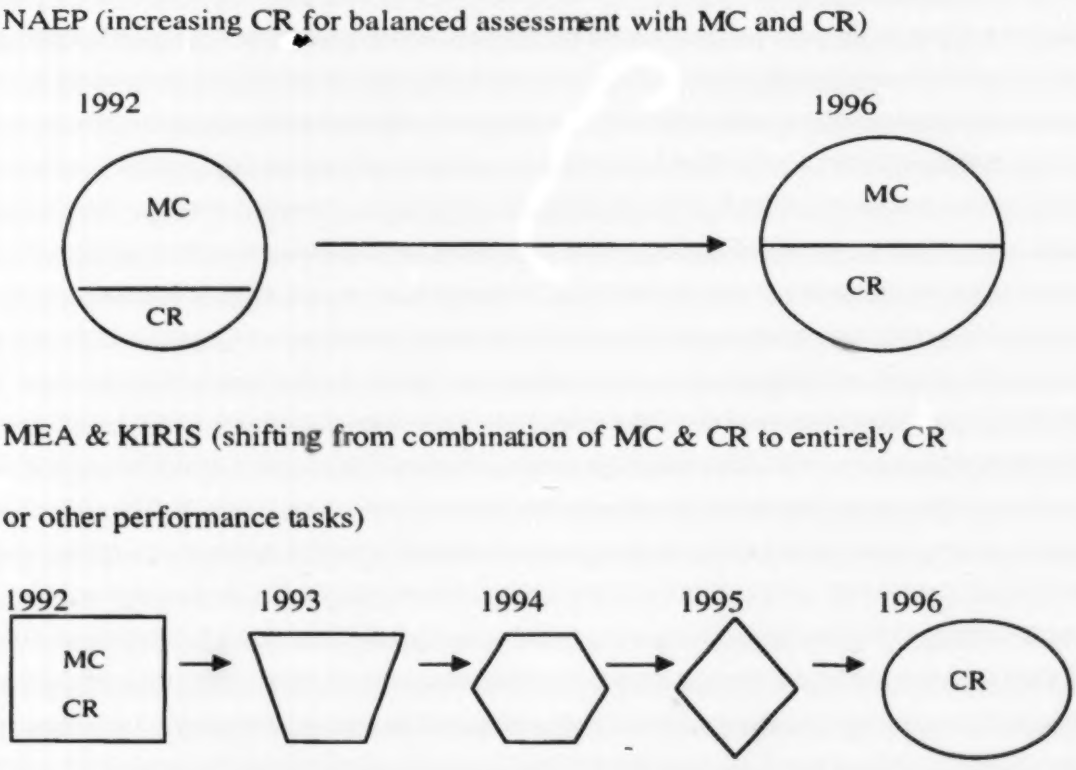

Figure 2. Changes in the Format of NAEP, MEA, and KIRIS from 1992 to 1996

Reliable estimation of achievement gains depends on robust test equating. NAEP, Kentucky and Maine assessments all used equivalent scaling and equating methods based on Item Response Theory. Nevertheless, there are differences between the NAEP and state assessments in their test equating frequency. NAEP equating was done directly between 1992 and 1996. MEA and KIRIS, which administer assessments every year, equating was done successively, that is, equating the 1993 assessment with its 1992 counterpart, the 1994 assessment with its 1993 counterpart and so on. Arrows in Figure 2 illustrate the difference in test equating process. This affects the reliability of equating: the equating of 1992 and 1996 test results is likely to be more reliable in NAEP than in the state assessments. In both the KIRIS and MEA relatively smaller percentages of items were used for equating, and this also might have increased the error of equating.

KIRIS proficiency level cut points for Accountability Cycle II (92/93 – 95/96) were linked to corresponding points for Cycle I (91/92 – 93/94). The method of linking was to determine the relationship between the original and revised 1992-93 scales using a linear transformation method (conversion of cut points based on changes in the mean and standard deviation of scale scores), and adjusting the proficiency level cutpoints accordingly. The accuracy of this adjustment also could have affected the gain in percent of students at the Advanced level from 1992 to 1996.

If equating happens regularly between successive years, the comparison of test results from remote years becomes less reliable because of the accumulation of equating errors. In other words, the link between 1992 and 1996 state assessment results should become more tenuous as a result of more drastic changes in the format of test as well as more frequent test administration and equating. To test this hypothesis, we attempted to check the stability of the linkage between the 1992 and 1996 state assessments by equating the two tests directly and comparing the results with the original gain scores that were obtained through the "chain-link" equating strategy. However, we found that there were no common items in the 1992 and 1996 MEA math assessments, which makes it impossible to equate them directly.

Differences in Test Stakes

In addition to the potential impact of changes in test format and related equating problems, one of the reasons for the greater achievement gains in Kentucky and Maine based on their state assessments might be the impact of the state assessments on school curriculum and instructional practices due to the stakes attached to the state test results. While there may be many other reasons for overstated or understated achievement gains (Wise & Hoffman, 2002), we here focus on the impact of high-stakes vs. low-stakes testing.

It is difficult to quantify how high the stakes of testing were and how much influence it might have had on actual test results. But when we simply compare the stakes of three assessments in terms of the consequences of testing for schools and school systems, it becomes obvious that the KIRIS has higher stakes than the MEA, which in turn has higher stakes than the NAEP (see Figure 3). In Kentucky, scores were used to measure school improvement and to give schools rewards or sanctions based on the adequacy of year-to-year progress. Not as high-stakes a test as the KIRIS, the MEA was designed primarily to provide information to schools to assist in making decisions about curricula and instruction. Reporting school performance to the public was also likely to produce moderate pressure on schools. This comparison of test stakes at the school or school district level, however, does not apply to the student level where neither state gave individual students substantial incentive to perform well on the state tests.

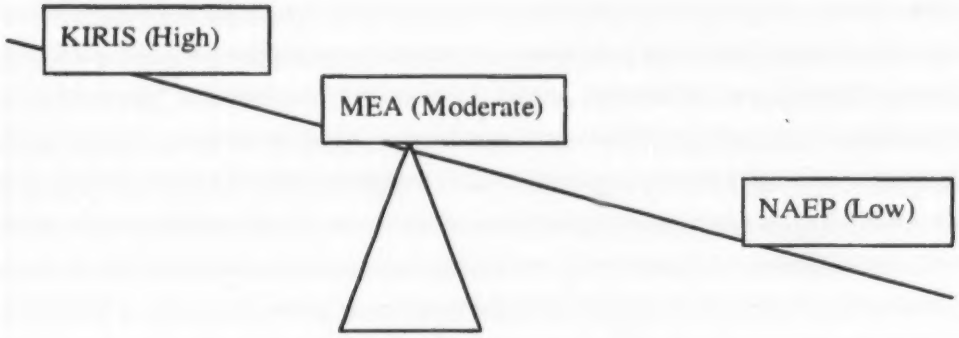

Figure 3. Contrast of NAEP, MEA, and KIRIS in the Level of Test Stakes

Given such moderate to high stakes attached to the KIRIS and the MEA for schools, it is likely that state assessment results show much greater improvement than national test results reveal. Linn (2000) explains the problem as follows:

> Divergence of trends (between a state's own assessment and NAEP) does not prove that NAEP is right and the state assessment is misleading, but it does raise important questions about the generalizability of gains reported on a state's own assessment, and hence about the validity of claims regarding student achievement. (p. 14)

The KIRIS technical manual noted that Kentucky students achieved gains on both NAEP and KIRIS but disregards the difference in the size of gains by saying that "As long as each measure provides an indication of whether changes over time are statistically significant, it is possibly to compare trends broadly. Comparing the magnitude of changes on one measure with magnitude of changes on another is more complicated, especially when multiple sets of scores are available for one or the other of the measures (such as scale score and standards-based percentage estimates) (KDE, 1997). But at the same time the manual raises the caution that some improvement in KIRIS scores is likely to occur as a result of directing school curricula toward the high-stakes test and preparing students for the test.

Our finding of the greater achievement gains in both Maine and Kentucky based on their own state assessments is consistent with the hypothesis that state assessments with serious consequences for schools would result in greater gains than NAEP without any stakes. However, our comparison of the two states in the amount of differences between NAEP and state assessment gains does not consistently support the expectation that Kentucky with relatively higher stakes would show greater differences than Maine; Maine reported greater gain than Kentucky at grade 4 while the pattern is reversed at the 8th grade level.

III. Discussion

Evaluation of systemic school reform requires us to investigate the adequacy and utility of the currently available data for assessing and understanding the performance of education systems. This study addressed two interrelated questions regarding the use of national and state assessment data. First, do national and state assessments provide the same information on the performance of states? Second, what are the factors that might explain the discrepancies between national and state assessment results? Kentucky and Maine were chosen for this study in which three key aspects of educational system performance were examined: achievement level, achievement gap, and achievement gain.

One might simply argue for using state assessment alone for evaluation of systemic school reform because it should be better able to capture the impact of state education reform policies than NAEP. It might be true that a state assessment better reflects state-specific reform goals because of stronger alignment with state curriculum standards, but it is also true that national assessment is more relevant to evaluating systemic reform that often goes beyond the boundary of a particular state given the influences of national standards and interstate benchmarking or comparisons. Table 19 provides a summary of consistent and inconsistent results in the national and state assessments as well as the factors that may account for the differences and should be considered in comparing and combining NAEP and state assessment results.

While there were seemingly close similarities between the four categories in NAEP and the corresponding four categories in state assessments, the percentage of students who perform at or above high proficiency levels in the Maine and Kentucky assessments (i.e., 'Advanced' on the MEA, 'Proficient' on the KIRIS) were not totally consistent with the national assessment results (i.e., 'Proficient' on the NAEP). Many other states also reported different results, but they tended to show the opposite patterns, i.e., greater percentage of students meeting the standard on the state's own assessment than on the NAEP. This indicates that these two states' assessment standards were uniquely higher than NAEP. However, the results were not entirely consistent across grades and years. This inconsistency might be due to differences between NAEP and state assessments in the definitions of performance standards and the methods of standards-setting. Therefore, extra caution is needed when comparing and/or combining the results on performance levels from NAEP and state assessments.

The national and state assessments were relatively consistent in their estimation of achievement gaps between students with different background characteristics. However, the size of achievement gaps were slightly smaller on the state assessments than on NAEP. Differences in the testing sample and the test itself may have influenced the results. While there was no significant difference between NAEP and state assessment data in the representation of major groups related to gender, race, and parental education, Title I students were not equally represented in the two assessments. On the other hand, NAEP had a wider range of item difficulty than the state assessments, and thus was better able to differentiate students performing at different achievement levels. These differences make it difficult to compare the size of the student achievement gaps between NAEP and state assessments. A further complicating factor is the possibility that state assessment had a greater impact on lower-performing students and their schools when they paid more attention to the state test as an accountability measure and teach to the test.

Both states reported increased student achievement based on their statewide assessment results. Because the NAEP and state assessments employed different scales for test scores, a common metric in standard

deviation units was established. The sizes of achievement gains from state assessments (i.e., gain scores from 1992 through 1996) turned out to be greater than their counterparts from NAEP. The state assessments went through more drastic changes in test format and more frequent test equating, which might have influenced the reliability of achievement gain estimates. Also, it is possible that student achievement gains were inflated by states' own assessments that were high-stakes tests and thus have had greater impacts on curriculum and instruction than NAEP.

This study explored a limited number of factors which might explain the discrepancies between national and state assessment results on school system performance. Further studies are needed to test not only the hypotheses presented in this report but also other alternative hypotheses. The findings from the two selected states may not be generalized to all states. With these caveats in mind, the study pinpoints the areas of consistency and inconsistency in the NAEP and state assessment results. It suggests that educational policymakers and practitioners become more aware of differences between current national and state assessments and potential biases and limitations in using only one of the two assessments to evaluate statewide educational system performance.

Table 19. Evaluation of the National (NAEP) and State (MEA/KIRIS) 4th and 8th Grade Math Assessment Results on Maine and Kentucky Education System Performance

	What the national and state assessments commonly say about	What the national and state assessments say differently about	What may account for the differences and should be considered for evaluation
Performance level	Majority of students found to perform below the Proficient/Advanced achievement level.	Percentage of students performing at or above Proficient level was smaller on state assessments than on NAEP. The size of this difference was also inconsistent across grade and year.	1. NAEP was more specific than KIRIS in defining its performance standards. 2. MEA standards were more rigorous than NAEP. 3. NAEP used test-centered standards-setting methods, whereas MEA and KIRIS used examinee-centered methods.
Achievement gap	The achievement gaps among different racial and socioeconomic groups of students were significant.	The achievement gaps were slightly smaller on state assessments than on NAEP. The size of this difference varied among the type of groups compared.	1. Percentage of Title I students in NAEP differed from its counterpart in MEA and KIRIS. 2. NAEP used test items with wider range of item difficulty than MEA and KIRIS.
Achievement gain	Statewide achievement gains (measured by increases in scale score or percent proficient) from 1992 to 1996 were positive.	The achievement gains were substantially smaller on state assessments than on NAEP.	1. MEA and KIRIS went through greater changes in test format and more frequent test equating than NAEP. 2. MEA and KIRIS had higher test stakes than NAEP.

References

American Educational Research Association, American Psychological Association, & National Council on Educational Measurement (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

Bond, L. A., Braskamp, D., & Roeber, E. R. (1996). *The status of state student assessment programs in the United States: Annual Report.* Oakbrook, IL: NCREL.

Consortium for Policy Research in Education (1995). *CPRE policy briefs. Tracking student achievement in science and math: The promise of state assessment systems.* New Brunswick, NJ: Rutguers University.

Dossey, J. A., Mullis, I. V. S., & Jones, C. O. (1993). *Can students do mathematical problem solving?: Results from constructed response questions in NAEP's 1992 mathematics assessment.* Washington, DC: OERI, U.S. Department of Education.

Hambleton, R.K., Jaeger, R.M., Koretz, D., Linn, R., Millman, J., & Phillips, S.E. (1995). *Review of the measurement quality of the Kentucky Instructional Results Information System, 1991-1994.* A report prepared for the Office of Educational Accountability, Kentucky General Assembly.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (pp. 485-514). New York: Macmillan.

Jaeger, R. M., & Mills, C, N. (2001). An integrated judgement procedure for setting standards on complex, large-scale assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 313-338). Mahwah, NJ: Lawrence Erlbaum Associates.

Kentucky Department of Education (1995). *KIRIS Accountability Cycle 1 Technical Manual.* Kentucky: Author.

Kentucky Department of Education (1997). *KIRIS Accountability Cycle 2 Technical Manual.* Kentucky: Author.

Laguarda, K. G. et al. (1994). *Assessment programs in the statewide systemic initiatives (SSI) states: Using student achievement data to evaluate the SSI.* Washington, DC: Policy Studies Associates.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher.* 2(29), 4-16.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district results to national norms: The validity of the claims that "everyone is above average." *Educational Measurement: Issues and Practice.* 9(3), 5-14.

Maine Department of Education (1996). *1994-95 MEA Performance Level Guide: Grade 4.* Maine: Author.

Maine Department of Education (1996). *1994-95 MEA Performance Level Guide: Grade 8.* Maine: Author.

National Center for Education Statistics (1997). *Technical report of the NAEP 1996 state assessment program in mathematics.* Washington, DC: OERI.

National Education Goals Panel (1996). *Profile of 1994-95 state assessment systems and reported results.* Washington, DC: Author.

National Research Council (1999). *Uncommon measures,* M. Feuer, P. W. Holland, B. F. Greeen, M. W. Bertenthal, & C. Hemphill (Eds.)., Committee on Equivalency and Linkage of Educational Tests. Washington, DC: National Academy Press.

Neil, M., Bursh, P., Schaeffer, B., Thall, C., Yohe, M., & Zappardino, P. (1995). *Implementing performance assessments: A guide to classroom, school, and system reform.* Cambridge, MA: FairTest.

Reese, C. M., Miller, K. E., Mazzeo, J., & Dossey, J. A. (1997). *NAEP 1996 mathematics report card for the nation and the states.* Washington, DC: OERI.

Shepard, L. A., Glaser, R., Linn, R. L., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement: An evaluation of the 1992 achievement levels* (A report of the National Academy of Education Panel on the evaluation of the NAEP trial state assessment). Stanford, CA: National Academy of Education.

U.S. General Accounting Office (1993). *Educational achievement standards: NAGB's approach yields misleading interpretations* [GAO/PEMD-93-12]. Washington, DC: Author.

Wise, L. L., & Hoffman, R. G. (2002). How will assessment data be used to document the impact of educational reform? In R. W. Lissitz & W. D. Schafer (Eds.) *Assessment in education reform: Both means and ends* (pp. 146-161). Boston: Allyn and Bacon.

TM034388

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

Reproduction Release
(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Using national and state assessments to evaluate the performance of state educatio	
Author(s): Jaekyung Lee and Walter McIntire	
Corporate Source:	Publication Date: July 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	**Level 2A**	**Level 2B**
☑	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: Jack Lee	Printed Name/Position/Title: Jaekyung Lee, Assistant Professor	
Organization/Address: University of Maine, 5766 Shibles Hall, Orono, ME 04469	Telephone:	Fax:
	E-mail Address: jklee@umit.maine.edu	Date: 7-30-02

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org